Joyce Choe

BIOS 669: Final project

Recalled Food Products in the U.S.

Abstract/Goals:

The Food Enforcement Report includes food products that have been recalled from the market because they do not meet standards set by the Food and Drug Administration (FDA). The laws that FDA regulate aim for consumer safety and public health. For this project, I will focus on food products distributed to North Carolina (NC) with a terminated recall status, which means that the recall investigation has been thoroughly done by correction or removal of the recalled product. The open FDA API (Application Programming Interface) will provide the enforcement report data from June 2012 to present. The primary goal is to display the frequency of recalled food products distributed to NC by firms from a specific food business center region. Another goal is to calculate the number of days between the recall initiation date and termination date to test and visualize relationships between number of days and region, reason for recall, initial firm notification, and voluntary-mandated status. Using these methods, exploratory data analysis will be demonstrated, so that these initial findings will be useful for statistical testing of where and when recalled food products distributed to NC are most likely to occur.

Data Sources:

The data for this project are publicly available:

1) openFDA food enforcement report API

2) USDA look-up infographic about food business center regions.

The first source includes information about individual recalled food products in the U.S. from 2004 to present (April 30, 2024). This API is updated weekly. The second data source called look-up information, is used to categorize location of food firms to a food business center region. For this project, I modified the look-up information to not include Intertribal region and to only include twelve regions: Northwest and Rocky Mountain, Southwest, North Central, Heartland, Rio Grande Colonias, Great Lakes Midwest, Delta, Appalachia, Northeast, Southeast, Islands and Remote Areas, and International. After merging these two data sources, I subset the data set for recall initiation dates between June, 1, 2012 – present, recalls with terminated status, and recalls distributed to NC. Also, I included only unique recall event IDs and excluded missing values from variables of interest.

Methods:

The key variables I used in this project are listed in Table 1. The variables with asterisks were derived from calculation or re-classification of existing variables. These variables will be used to visualize trends and fit statistical tests to summarize the data set. For visualization, I plotted series and stacked bar plots to show frequency of recalls each year by region. Then, I plotted multiple boxplots showing number of days to process a recall to completion (response variable) by several categorical variables. The six categorical variables chosen for the boxplots include voluntary mandated status, region, general reason, initial firm notification, classification, and year. For each group within a set of boxplots, I displayed the medians and number of observations because the observations were positively skewed. Next, assuming independence of recalled food events and observing a non-parametric distribution of the response variable, I performed multiple Kruskal-Wallis tests on the relationship between days and the same six categorical variables. Kruskal-Wallis tests whether at least one median between levels of a categorical variable is significantly different from another level, while the null hypothesis is that the median levels are equal.

Results/Findings:

Exploratory data analysis shows that the number of recall events were surprisingly low when examining categorical variables: voluntary mandated status and geographic food business center region. For example, there were a total of 7 FDA-mandated recalls and 1879 voluntary: firm initiated recalls from June 2012 to present. Additionally, International region had a total of 3 recalled food products compared to 23 recalled food products in Island & Remote areas and Delta (mainland) in the same time period.

Since discrepancies in frequency of recalls were present for each categorical variable, the median number of recalls for each level were compared using Kruskal-Wallis test. Kruskal-Wallis test showed that the number of days to process recalls differed significantly by year. Year 2015 had the highest median days (344), while 2024 has the lowest median days (69) as expected, since the data for 2024 exists up to April 30, 2024. Surprisingly, 2012 was relatively high even though the earliest data point begins on June 2012.

The Kruskal-Wallis test also showed at least one significant difference in the number of recalls by region. The international region had the highest median days (497) to process recall events, although with a limited sample size of 3, while the Southwest region had the lowest median days (132). All other categorical variables, voluntary status, reason, notification, and classification did not significantly differ in median days. For nearly all of the grouped variables, the boxplots were positively skewed, which makes sense since we would hope for or expect most recalls in the U.S. to be processed quickly. These are a few highlights from this exploratory data analysis. In conclusion, descriptive statistics and exploratory data analysis facilitate recognition of patterns in the data that would otherwise not have been known.

Limitations

The raw data set was derived from the openAPI food enforcement site, where data were noticeably ‘untidy’ prior to analysis. For example, several values in the data set were missing or mis-typed under the reason\_for\_recall variable. Thus, I parsed the text to generalize each unique recall to one of six broad categories: Unprepared, Precaution, Microbe, Mislabeled, Contaminant, and Other. Examples of ‘Unprepared’ include undercooked, under washed, or any other means of human labor where the food was not prepared safely. ‘Precaution’ category represents food products that are potentially unsafe and recalled even if the hazard has not been confirmed, to prevent risk of consuming potentially hazardous food. ‘Microbe’ covers confirmed outbreak or known pathogens infecting the food product resulting in foodborne illness or a public health issue. ‘Contaminant’ includes any source of inadvertent or undesirable contamination that is not microbial in nature. The ‘Other’ category includes all remaining reasons that do not belong to the first five categories.

Table 1. Key variables used in this project

|  |  |
| --- | --- |
| **Name** | **Description** |
| classification | Class I (adverse), Class II (somewhat adverse), Class III (unlikely adverse) |
| \*date1year | Year of recall initiation date; calculated as substring from recall\_initiation\_date variable |
| \*days | Number of days between termination-initiation dates; calculated difference between dates |
| distribution\_pattern | Places in the U.S. where recalled food products were distributed |
| \*general\_reason | General reason for recall; parsed text from reason\_for\_recall variable to generalize reasons |
| initial\_firm\_notification | Recall notification method |
| reason\_for\_recall | Statements on reason for recall |
| recall\_initation\_date | Date when recall was first notified |
| recall\_termination\_date | Date when recall investigation is done |
| state | State location of firm |
| \*usda\_region | Food regional business center; classified state to a usda\_region |
| voluntary\_mandated status | Recall was voluntarily reported by firm or mandated by FDA |

References

API:

<https://open.fda.gov/apis/food/enforcement/>

<https://api.fda.gov/food/enforcement.json>?

Look up table reference state by region:

<https://www.ams.usda.gov/services/local-regional/rfbcp>

Content:

<https://www.fda.gov/safety/recalls-market-withdrawals-safety-alerts/enforcement-reports>

<https://www.fda.gov/safety/enforcement-reports/enforcement-report-information-and-definitions#report_label>